Preprint
UCRL-JC-144382

Interpreting Microarray Data to Build Models of Microbial Genetic Regulation Networks

B. A. Sokhansanj, J. B. Garnham, J. P. Fitch

This article was submitted to
Photonics West 2002, San Jose, CA, January 19-25, 2002

January 23, 2002

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

This work was performed under the auspices of the United States Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

Interpreting microarray data to build models of microbial genetic regulation networks

Bahrad A. Sokhansanj*, Janine B. Garnham, J. Patrick Fitch

Biology & Biotechnology Research Program, Lawrence Livermore National Laboratory, University of California, Livermore, CA, USA, 94550

ABSTRACT

Microarrays and DNA chips are an efficient, high-throughput technology for measuring temporal changes in the expression of message RNA (mRNA) from thousands of genes (often the entire genome of an organism) in a single experiment. A crucial drawback of microarray experiments is that results are inherently qualitative: data are generally neither quantatively repeatable, nor may microarray spot intensities be calibrated to *in vivo* mRNA concentrations. Nevertheless, microarrays represent by the far the cheapest and fastest way to obtain information about a cell s global genetic regulatory networks. Besides poor signal characteristics, the massive number of data produced by microarray experiments poses challenges for visualization, interpretation and model building. Towards initial model development, we have developed a Java tool for visualizing the spatial organization of gene expression in bacteria. We are also developing an approach to inferring and testing qualitative fuzzy logic models of gene regulation using microarray data. Because we are developing and testing qualitative hypotheses that do not require quantitative precision, our statistical evaluation of experimental data is limited to checking for validity and consistency. Our goals are to maximize the impact of inexpensive microarray technology, bearing in mind that biological models and hypotheses are typically qualitative.

Keywords: microarrays, bacteria, gene regulation, fuzzy logic, modeling and simulation

1. INTRODUCTION

Technological advances in DNA sequencing have led to the complete sequencing of the human genome (approx. 30,000 genes) and the feasibility of sequencing an entire microbial genome (approx. 500-5000 genes) in a matter of weeks or even days. However, the sequence of genes is only the parts list for the cell. Cellular function arises from the interaction of proteins and regulatory sequences in DNA governed by the response to environmental stimuli and signals from other cells, with multiple feedback as illustrated in Fig. 1.

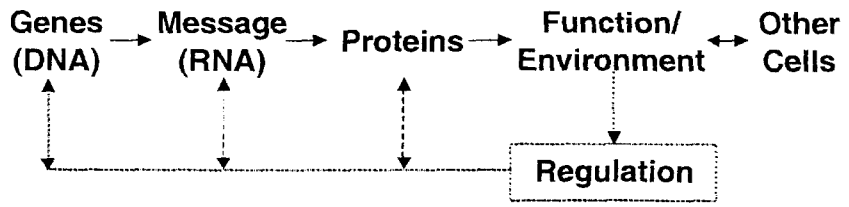


Figure 1: Coding sections of the DNA sequence (genes) are expressed : transcribed to message RNA (mRNA). The mRNA is then translated to proteins that catalyze metabolic reactions and form cellular structures, determining cell function. Some *regulatory proteins* interact with regulatory sections of DNA to control the transcription rate of mRNA; in general, regulatory feedback can occur at any step of protein production.

Large-scale biology is now in its Post-Sequencing era, characterized by engineering advances [1] towards temporal profiling of gene and protein expression, high-throughput protein structure prediction and determination, and identification of protein-protein interactions. Transcription profiling techniques, such as photolithographic Affymetrix

* sokhansanj@llnl.gov; phone 1 925 422-8643; fax 1 925 422-2282; Lawrence Livermore National Laboratory, L-452, 7000 East Ave., Livermore, CA, USA 94550

DNA chips[2] and spotted glass microarrays with optical fluorescence detection[3], allow the simultaneous measurement of mRNA expression from each genes in a cell s complete genome. A typical two-sample competitive hybridization microarray experiment is described in the schematic shown in Fig. 2. After fluorescence images are analyzed, the usual result of a microarray experiment is the relative expression of each gene: the ratio of fluorescent intensity under one experimental stimulus to intensity under the other stimulus (or control state)[4]. Taking a ratio magnifies already considerable sources of experimental uncertainty. Microarray fabrication is often inconsistent: the size, shape, and alignment of spots varies from array to array, resulting in poor repeatability. Defects in the slide or incomplete washing, can result in local regions of higher background. Finally, results are highly sensitive to methods of image analysis, such as spot recognition, intensity quantification, dye intensity normalization, and background subtraction[5]. There are more consistent fabrication and analysis methods for DNA chips; however, they are still sensitive to experimental protocols and are considerably more expensive than glass microarrays.

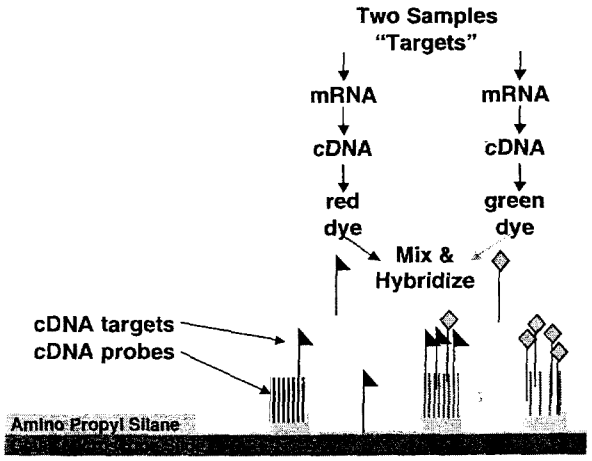


Figure 2: DNA microarray experiment measuring changes in gene expression. Message RNA is isolated from two samples of cells exposed to a different set of conditions, e.g. temperatures of 26¡C and 37¡C. The mRNA from each sample is separately labeled with two distinct fluorescent dyes (by PCR to labeled cDNA). The array is spotted with single stranded DNA corresponding to each gene in the genome. Sample targets are added to the array, where they hybridize with complementary cDNA probes. After washing, the fluorescence signal of each dye measured at a spot corresponding to a particular gene is proportional to the amount of mRNA produced by that gene in the respective sample.

Even with all the well-known drawbacks of microarray experiments, they remain very popular. Microarrays are relatively fast and cheap, and arrays are currently the only practical method for whole genome expression studies required to identify the global regulatory networks of the cell. There has been considerable recent work to improve quantification of microarray results and apply novel statistical methods, e.g. to replace lost data, filter noise, study the effect of replication, etc.[6]. However, the advent of RT-PCR, which can provide a rapid and accurate assay for a hundred genes at a time, allows scientists to confirm the behavior of genes of interest found in microarray experiments. Thus, it seems that microarray experiments are most useful to identify global expression patterns, build and test initial rough draft system models, and screen for genes and subpathways to study more closely with more accurate methods and knockout experiments. The two most important goals are determining whether past models of regulation remain consistent with genome-scale data and identifying unexpected expression patterns. Biological models are typically linguistic and/or graphical: behavior is described using phrases like weakly inhibited or strongly expressed . Fuzzy logic models based on linguistic rules[8] have been successfully used to practically model control systems that are structurally similar to biological pathways. Working within the limitations of typical microarray data quality, we are developing a framework for qualitative modeling using fuzzy logic to help achieve the goal of deriving maximum value from microarrays.

Before biologists can begin building models and screening for genes of interest, it is necessary to organize the massive amount of data from microarray experiments (thousands of data points for each gene, frequently taken at several time points). The most popular approach has been to cluster genes according to similar temporal profiles of expression[7]. Several methods have been proposed; however, all are in question due to the complexity of genome regulation in eukaryotic cells (cells with a nucleus, e.g. yeast, mice, human cells). However, while gene regulation in bacteria (prokaryotic cells that lack a nucleus) is not yet perfectly understood, regulatory mechanisms are much less diverse and better understood. Microarrays are particular useful for studying cell function of bacteria. Unlike in eukaryotic cells, mRNA transcript in bacteria is very unstable and proteins are relatively short-lived, thus protein production in bacteria closely tracks transient changes in mRNA transcription. Thus, microarrays give a fairly good indication of the protein activity of the bacteria as well, making arrays even more comprehensive and useful for studies of bacterial regulation. Analysis of transcription patterns is also easier for bacteria than eukaryotic cells because genes are generally regulated in simple linear blocks. The block structure of transcriptional regulation is illustrated in Fig. 3 for the *lac* operon of *E. coli*, the most studied system of gene regulation. The transcription of several genes in a consecutive sequence (*operon*) along the chromosome occurs simultaneously with common regulatory sites. In some cases, short regulatory sequences between genes can attenuate or enhance the transcription of some genes in the operon. We have developed a Java tool to help identify operons and their common regulatory features from microarray experiments.

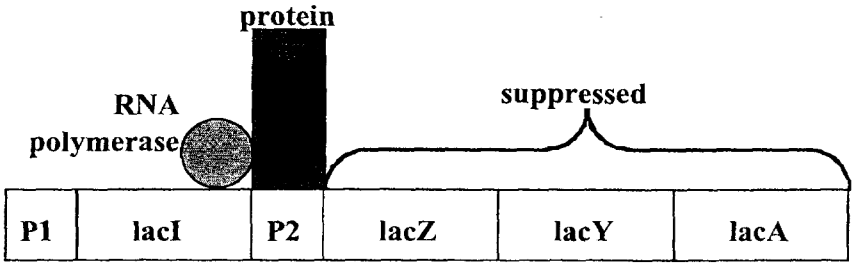


Figure 3: Part of the *E. coli* chromosome coding for the enzymes responsible for lactose metabolism (proteins coded by genes lacZ and lacY). The DNA sequence includes promoter regions where RNA polymerase (RNAP) binds to begin transcribing the genes in the operon. At the end of the operon is a terminator region where the RNAP detaches. The lacI gene produces a protein that normally binds to the promoter of lacZ and lacY (P2), blocking their transcription by RNAP. When lactose is present, it binds to lacI allowing the RNAP molecule to continue down the operon and produce enzymes to consume the lactose. Once the lactose is gone, lacI again binds to P2 and lacZ and lacY transcription ceases, thus the lac operon is a negative feedback system.

2. MICROARRAY DATA ANALYSIS

In our laboratory, we are using microarrays to reveal temporal profile global regulation of the *Yersinia pestis* low calcium response . Virulence factors coded on the plasmids of *Y. pestis* , the bacteria that cause plague, are induced *in vitro* when the temperature increases to 37¡C from 26¡C and Ca^{2+} is absent, mimicking conditions inside human tissue during infection. There is evidence the *Y. pestis* chromosome also contains genes involved in virulence, and those genes are also controlled by temperature and Ca^{2+} concentration changes. The goals of our microarray experiments are to verify previous models of virulence regulation, screen for previously unknown genes and operons involved in low calcium response, and ideally build the rough drafts of new regulatory models.

Accurate measurement of transcription is *not* an objective of our experiments. Key discoveries will be verified using more accurate (and lower throughput) technologies. Thus, our preliminary data analysis has the primary objective of ensuring as many interesting signals are found as possible, and reducing the number of false positives (which increases the required number of costly verification experiments). At the same time, we can not afford large numbers of replicant experiments, and thus can not use many of the sophisticated statistical tools discussed in recent literature. There are repeated spots on each chip to test variation between spots. However, there are no repeated chips: the variation from chip to chip is often so high that as many as ten replicant slides would be required to achieve statistical significance. Rather, we try to establish the consistency of data in two ways. For each time point, we measure six slides: the ratio of each combination (of 26¡C/No Ca^{2+}, 37¡C/No Ca^{2+}, and 37¡C/w. Ca^{2+}) and each ratio with dye colors reversed.

Fig. 4 illustrates how differences in dye binding characteristics and background require data normalization. We do not subtract for background, as recent studies have shown that unhybridized spots typically have very low intensities. Consequently, the most accurate background values are lower than those given by software (which usually takes into account local background around the spot). Dye intensities are normalized by dividing by the ratio of median intensities. The median is used because the mean is influenced by long tails resulting from transcription induced by the stimulus. In principle, the majority of genes should be unregulated in response to a specific stimulus, with a transcription ratio of approximately one. This assumption is not necessarily true in cases where the stimulus could cause a broad change in the transcription of many genes: this is observed in studies of cell cycle, starvation, etc. Temperature increase generally increases the metabolic rate and thus perhaps transcription as well, and the normalization will likely smooth out that effect. However, we would like to isolate and identify *genetic* mechanisms of thermoregulation, as opposed to any broad increase in transcription due to increased metabolism.

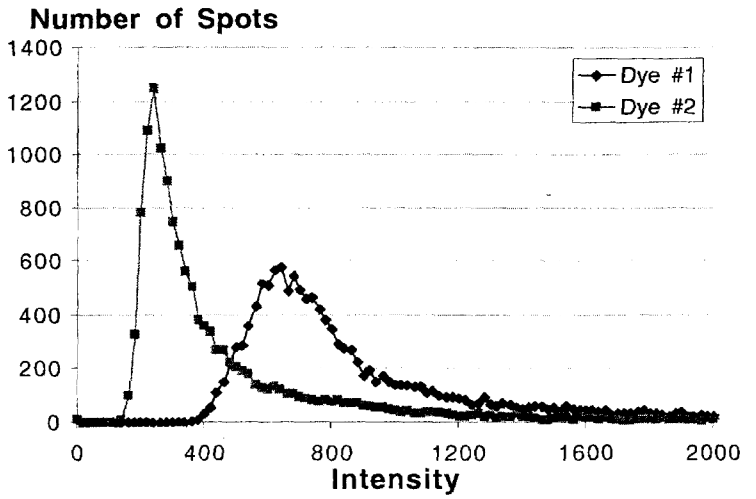


Figure 4: Intensity histograms for two dyes [V. L. Motin, unpublished data].

Once expression ratios have been obtained, we then apply a variety of methods to identify biologically interesting signals: standard deviations from the mean and other threshholds[9], Bayesian methods[10], and a plethora of methods developed by the statistics community[11]. Gene expression is also characterized in the context of operons, as discussed below. However, the statistical significance of our results suffers considerably due to the low number of replicants. As an additional verification of the significance of a gene expression ratio, we apply its closure relationship . Given 3 experimental conditions (i.e. A, B, and C) measured in all three combinations, the product of expression ratios for a particular gene should be unity, i.e.

$$\frac{A}{B}\cdot\frac{B}{C}\cdot\frac{C}{A}=1$$

or, using the (base 2) log of expression ratios that is usually reported,

$$\log\left(\frac{A}{B}\right)+\log\left(\frac{B}{C}\right)+\log\left(\frac{C}{A}\right)+\log(\phi)=0$$

where ϕ is a term representing a closure error phase, representing the relative error between the experiments. Fig. 5 shows a histogram of $\log(\phi)$ for genes in the set of six microarray experiments performed at a given time point. A perfectly consistent result has a $\log(\phi)$ of zero. Each ratio is taken as the average of the two reverse color experiments. This measure of consistency is used in conjunction with replicants to evaluate the risk that a gene expression signal is a false positive.

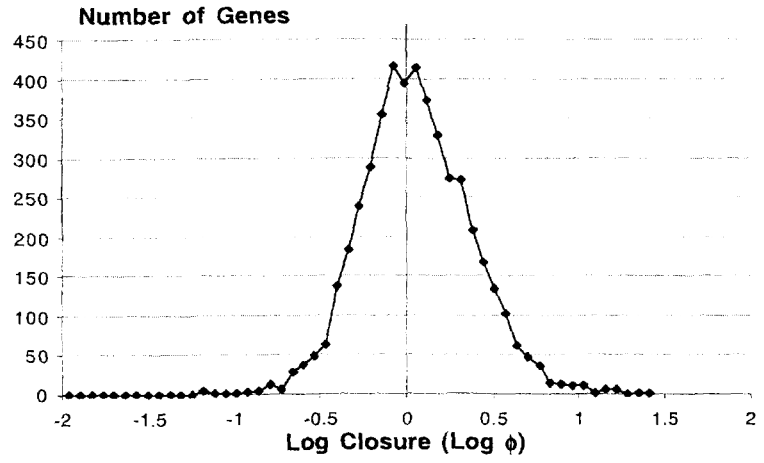


Figure 5: Histogram of closure error phases, log(ϕ) representing the relative error between experiments taken at different conditions.

3. OPERON IDENTIFICATION

Bacterial genes are spatially organized in operons as shown in Fig. 3 (though some operons only contain a single gene with its associated promoter). When gene expression is plotted in the spatial order of genes, correlation in gene expression signifies the potential for a continuous operon, illustrated in Fig. 6. This result suggests that we can apply several methods available to automatically recognize hypothetical operons, and these may be coupled with efforts to computationally predict operons based on sequence similarity and motifs[13].

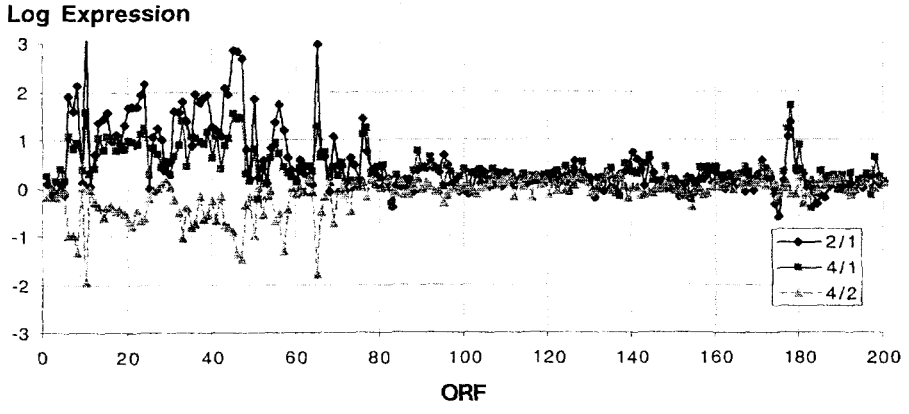


Figure 6: Gene expression in linear sequence along the plasmids of *Y. pestis*. Three expression ratios are shown, corresponding to combinations of the three experimental stimuli described previously. Correlated expression signifies a potential operon.

While automatic signal detection has an important role in locating features of large microarray data sets, there is no replacement for biological intuition. An important priority of our work is to allow biologists to visually navigate gene expression data and use their intuition to locate patterns of potential interest. Towards this goal, we have developed a Java applet that can be used to display the results of microarray time series for genes in circular chromosomes and plasmids. Screen shots of the Gene Viewer software are shown in Fig. 7. Features of the software include the ability to modify display scales and colors, zoom into regions of interest, simultaneously display multiple experiments, animate time series, and display annotation and sequence. In future versions, we plan to integrate the Gene Viewer with fuzzy models of regulatory pathways.

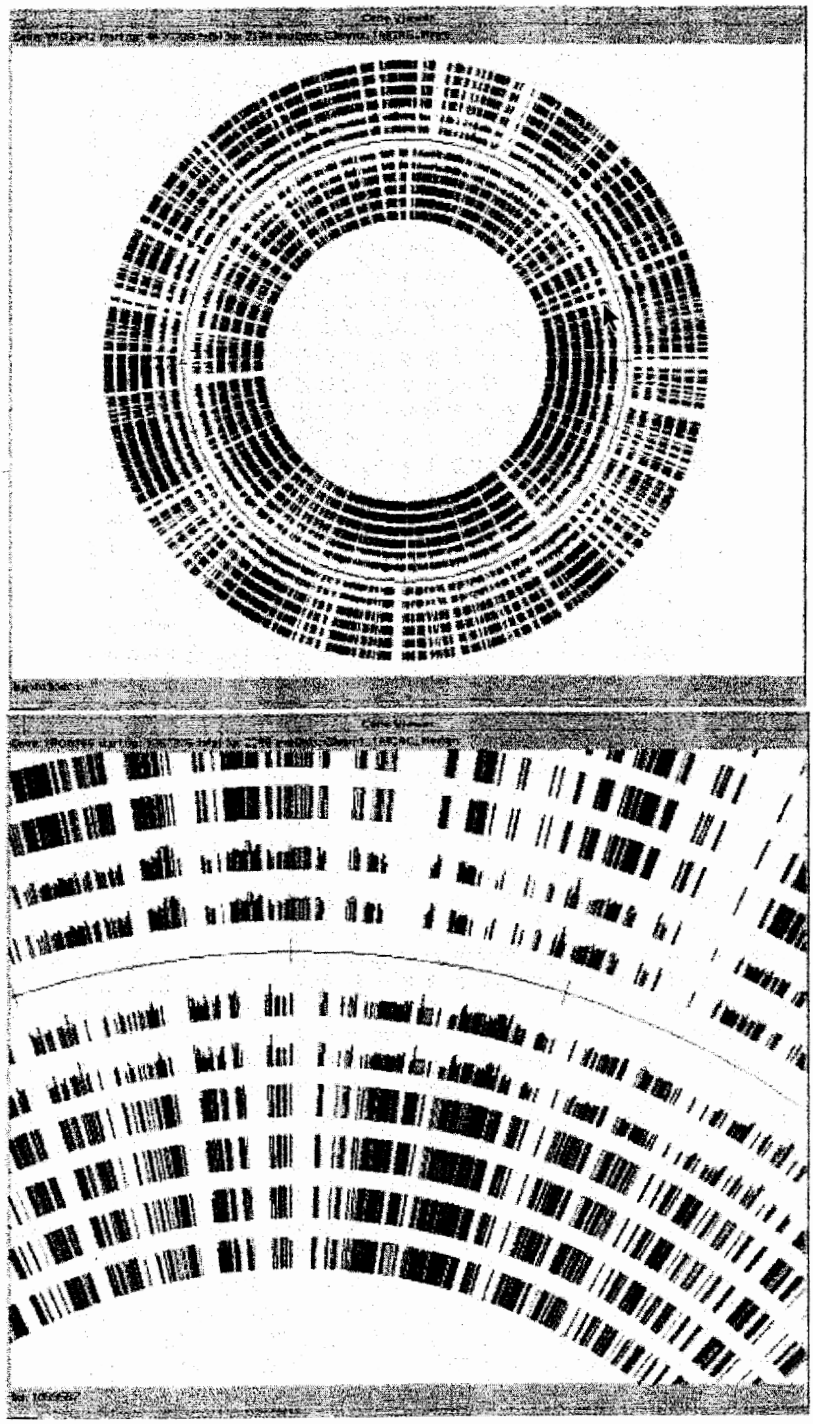


Figure 7: Screen shots of Gene Viewer software. Expression of the main chromosome of *Y. pestis* is displayed (approx. 4,000 genes), with six data sets shown. Each bar (see the bottom image for detail) is the expression of a gene. The angular position of each gene is exact. If a gene is coded in the forward direction, it is above the reference line, if it is in the reverse direction it is shown below the reference line. Green and red represent positive and negtive log ratios, while the height of the bars is proportional to expression level.

4. FUZZY REGULATORY MODELS

Typically, biologists qualitatively model a regulatory system, describing it in text or a diagram (e.g. see the schematic of the *E. coli* lac operon and accompanying caption of Fig. 9). Experimental results in biology are usually not quantitative (e.g. microarray data). While data may be analyzed quantitatively (e.g. enzyme assay), the intended conclusion is usually either qualitative (more or less) or acceptance/rejection of a hypothesis. Thus, experimental work is focused on making sure data are internally consistent and allow statistically valid conclusions, not generating accurate parameters for modeling the system chemically. As systems become more complex, qualitatively mental models become harder to develop and interpret. For complex biological systems as a system of equations (requiring accurate parameters) or as a binary logic network (on/off models are too simple to model many interactions). We propose that *fuzzy logic* is a natural language for modeling biology.

4.1 Fuzzy Modeling

A detailed discussion of fuzzy logic and modeling is outside the scope of this paper (see reviews elsewhere[15-17]). Briefly, fuzzy set theory is a generalization of classical set theory in which set membership becomes a degree continuously varying from 0 to 1.0 (where 0 represents absolute non-membership or false , and 1.0 corresponds to classical set membership or true). Fuzzy sets can be used to represent quantities, e.g. LOW, MED, and HIGH protein levels in a cell. Fig. 8 shows how quantitative values can be *fuzzified*, translated into a union of fuzzy set membership values. The inverse operation (*defuzzification*) is also possible, fuzzy logic models can be used in combination with quantitative data and models.

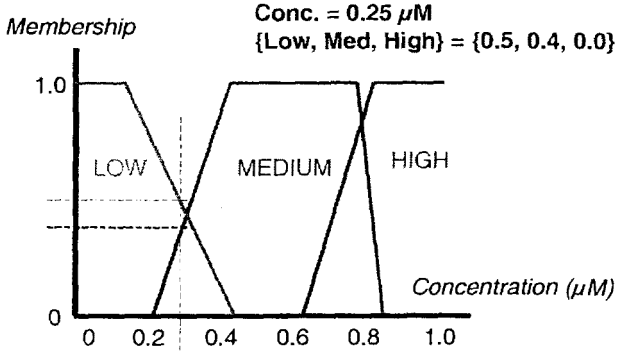


Figure 8: Fuzzification of a concentration (0.25 M) given a graphical definition of fuzzy concentration categories (Low , Med , and High).

Fuzzy versions of logical operators like AND and OR can be defined in many different ways. The simplest definition is to take the minimum of fuzzy set memberships for AND and maximum for OR. For example, given concentrations **C1** = {0.5, 0.4, 0.0} and **C2** = {0.1, 0.6, 0.2}; **C1** AND **C2** = {0.1, 0.4, 0.0}, C1 OR C2 = {0.5, 0.6, 0.2}. Another definition is product of memberships for AND and sum for OR, i.e. **C1** AND **C2** = {0.05, 0.25, 0.0}. A fuzzy rule typically takes the form IF **C1** AND **C2** THEN **Q** . In the simple variants of fuzzy logic we use in our models (min/max, sum/product), IF/THEN is synonymous with AND. In general, each node of a fuzzy model must have a rule accounting for every possible input state. For example,

IF **C1** AND **C2** THEN **Q**

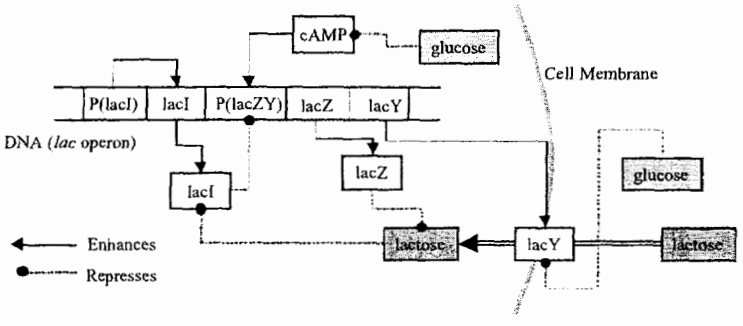
where C1 and C2 have 3 categories (Low, Med, High) there will be $3^2 = 9$ rules, e.g.

IF (**C1** = Low) AND (**C2** = Low) THEN (**Q** = Slow),
IF (**C1** = Low) AND (**C2** = Med) THEN (**Q** = Slow), . . .
. . .
IF (**C1** = Med) AND (**C2** = High) THEN (**Q** = Fast), . . .
. . .
IF (**C1** = High) AND (**C2** = High) THEN (**Q** = Fast).

If our model includes a node for the production level of a protein that depends on the strength of two promoters, translation rate, temperature, and 5 regulatory proteins (total of 9 inputs), that node alone would require $5^9 = 1{,}953{,}125$ rules! This curse of dimensionality has sharply limited the general use of fuzzy models for complex systems: most solutions (e.g. clustering, redundant rule elimination, etc.) are problem-specific and require significant domain knowledge. However, recently Combs has proposed an alternative format for fuzzy models called the Union Rule Configuration (URC)[17-18]. With this configuration, the above node would be rewritten as:

IF (**C1** = Low) THEN (**Q1** = Slow), IF (**C1** = Med) THEN (**Q1** = Fast),
IF (**C1** = High) THEN (**Q1** = Fast), IF (**C2** = Low) THEN (**Q2** = Slow),
IF (**C2** = Med) THEN (**Q2** = Med), IF (**C2** = High) THEN (**Q2** = Fast),
Q = **Q1** OR **Q2**.

Now, the node only contains 3*2 = 6 rules. With the URC, instead of growing exponentially with the number of inputs or fuzzy states, the complexity of the node grows linearly. Now, our example of the 9 input protein production model requires only 5*9 = 45 rules, and URC fuzzy logic becomes a computationally feasible modeling framework for complex biological systems. Fig. 9 shows the URC fuzzy model for the lactose regulation system of *E. coli*. In this case, the model includes both the production rates of proteins and their activity levels, along with promoter efficiencies. Quantitative data for the concentration of extracellular glucose and lactose may be fuzzification for use in the model. Fuzzy values of the model variables may be defuzzified and compared to quantitative data, or used as a fuzzy data point for comparison with qualitative data (e.g. result of a microarray experiment).



	lacI production				
P(lacI) strength	VL	LO	ME	HI	VH
	lacI activity				
lacI production	VL	LO	ME	HI	VH
lactose (in cell)	VH	ME	LO	LO	VL
protease	VH	HI	HI	LO	VL
	cAMP activity				
glucose	VH	HI	ME	LO	VL
	lacY, lacZ production				
P(lacZY) strength	VL	LO	ME	HI	VH
lacI activity	VH	HI	LO	LO	VL
cAMP activity	VL	LO	HI	HI	VH
	lacY activity				
lacY production	VL	LO	ME	HI	VH
glucose	VH	HI	ME	LO	LO
protease	VH	HI	HI	LO	VL
	lacZ activity				
lacZ production	VL	LO	ME	HI	VH
protease	VH	HI	HI	LO	VL
	lactose (in cell)				
lactose (outside cell)	VL	LO	ME	HI	VH
lacY activity	VL	LO	ME	HI	VH
lacZ activity	VH	HI	ME	LO	VL

Figure 9: Schematic and fuzzy models of *E. coli* lac operon regulation, showing enzymes (lacZ and lacY), substrates (glucose and lactose), regulatory proteins (cAMP and lacI), and promoters (P(lacI) and P(lacZY)) and genes in the linear sequence found on the chromosome. Lac operon is regulated by two kinds of negative feedback. One feedback loop is direct repression in absence of lactose, mediated by lacI interaction shown in Fig. 3. The other feedback loop, repression in abundance of glucose, is regulated by CRP-cAMP (shown as cAMP in the schematic), which enhances the promoter of lacZ and lacY. Glucose inhibits the activity of CRP-cAMP, thus reducing production of lacZ/Y.

4.2 Comparing Models to Data from Microarray Experiments

One important criticism of fuzzy physical models is the absence of statistically rigorous means of comparing predictions with experimental data and evaluating the performance of alternative models. However, it has already been shown that rigorous statements about microarray experiments are difficult to make. Data is of poor quality and not repeatable: if microarrays are to be used as screening or rough draft , strict rigor is much less important than rapid and intuitive computation. To compare microarray data with fuzzy predictions, expression ratios are fuzzified. Fig. 10 shows a simple scheme that is equivalent to a heavily smoothing linear interpolation. Results of all experiments are fuzzified using the same scheme, defined for the experiment with the largest dynamic range of ratios.

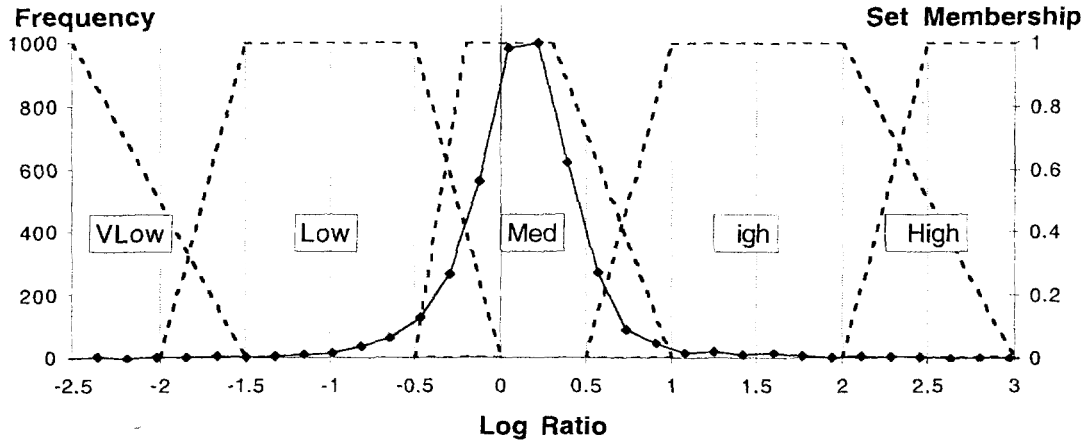


Figure 10: The fuzzification scheme for a set of microarray data is shown (dashed lines) for categories Very Low , Lower , Med (approx. no change), Higher , Very High . It is based on the distribution of log ratios (solid line with diamonds) from a single reference experiment (the one with the largest range of ratios).

The prediction of the fuzzy model is then cast in terms of relative expression, for example, expression with high Ca^{2+} is Lower than expression at lower Ca^{2+} . In general, the model prediction will be a union of fuzzy sets. Fig. 11. shows the simplest possible scheme to compare a single experimental data point to its predicted expression ratio. Model accuracy is defined by the maximum membership of the data point in the predicted fuzzy set. It corresponds loosely to the concept of possibility defined in fuzzy set theory, an analogy to classical probability[19]: the possibility that the prediction is true given the data point.

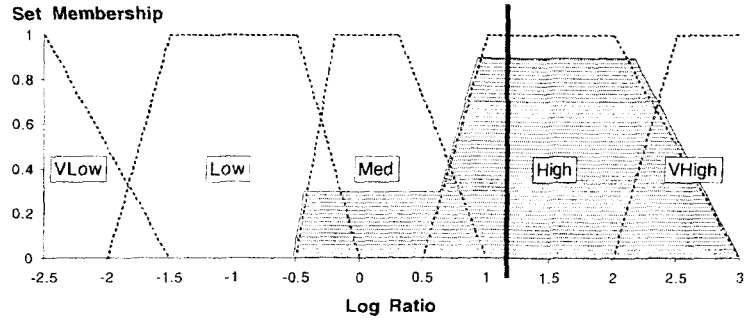


Figure 11: Possibility measure a model prediction of {0.3 Med, 0.8 High, 0.0 all other sets} given a gene with a log ratio of 1.2, based on the fuzzification scheme of Fig. 10. The resulting possibility (and metric for the prediction) is 0.8.

Increasingly, microarrays are being used to explore novel organisms and environmental responses with no previously described genetic regulation mechanisms. Even for biological problems with some prior conventional experimental investigation, increasing the number of time points and genes being studied by large-scale experiments makes the use of intuition less and less feasible. Consequently, an active area of current research is methods to infer models from complex data sets, i.e. microarrays and DNA chips. Boolean logic models are limited by a critical lack of resolution, and inverse problem solving with differential equations is insufficiently scalable. Previous attempts to learn fuzzy logic models from microarray data[19] were limited to only considering interactions between up to 2 factors, largely because of the curse of dimensionality. URC fuzzy models have a linear relationship between complexity and problem size and scale, thus the URC is an ideal framework for inferring models directly from data. Experience has shown that model learning problems previously on the scale of supercomputers can now be done using a PC in a few minutes[20], particularly with a

computationally trivial method for evaluating and comparing models, such as Fig. 11 describes. Approaches to building a model directly from data include adopting neural network learning and optimization algorithms (the URC is conceptually very similar to a perceptron or neural network model). Ultimately, our goal is to derive maximum value from some of the cheapest and easiest experiments available in a contemporary genetics laboratory. In the process we try to minimize the required accuracy and repeated experiments to ensure microarrays remain cheap, easy, and convenient tools for biologists to gain insight and gather intuition about the complex systems they study.

ACKNOWLEDGEMENTS

We thank our experimental collaborators at Livermore, Drs. V. L. Motin and E. Garcia for continued discussions and access to data. We also thank our colleagues L. Mascio-Kegelmeyer and D. Nelson for consultation about microarray image analysis and statistical analysis. BAS is supported by a supported by a LLNL Student Employee Graduate Research Fellowship. This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

REFERENCES

1. J. P. Fitch and B. Sokhansanj, Genomic engineering: moving beyond DNA sequence to function, *Proc. IEEE*, **88**, pp. 1949-1971, 2000.
2. D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobyashi, H. Horton, and E. L. Brown, Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nat. Biotechnol.*, **14**, 1675-1680, 1996.
3. J. L. DeRisi, V. R. Iyer, and P. O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, **278**, pp. 680-686, 1997.
4. Y. Chen, E. R. Dougherty, and M.L. Bittner, Ratio-based decisions and the quantitative analysis of cDNA microarray images, *J. Biomedical Optics*, **2**, pp. 364-374, 1997.
5. H. Hanspeter, D. Beule, S. Kielbasa, J. Korbel, C. Sers, A. Malik, H. Eickhoff, H. Lehrach, J. Schuchhardt, Extracting information from cDNA arrays, *Chaos* , **11**, pp. 98-107, 2001.
6. G. C. Tseng, O. Min-Kyu, L. Rohlin, J. C. Liao, and W. H. Wong, Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Res.*, **29**, pp. 2549-2557, 2001.
7. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, Cluster analysis and display of genome-wide expresion patterns, *Proc. Natl. Acad. Sci. USA* , **95**, pp. 14863-14868, 1998.
8. L. A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, *IEEE Trans. Syst. Man. Cybernet.*, **3**, pp. 28-44, 1973.
9. J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, Use of a cDNA microarray to analyse gene expression patterns in human cancer, *Nat. Genet.*, **14**, pp. 457-460.
10. T. Ideker, V. Thorsson, A. F. Siegel, L. E. Hood, Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data, *J. Comp. Biol.* , **7**, pp. 805-817, 2000.
11. Statistical analysis of gene expression from cDNA microarray experiments, http://www.statsci.org/micrarra/. Laboratory for the statistical analysis of microarrays, http://www-stat.stanford.edu/~tibs/lab/index.html. Terry Speed s microarray data analysis group, http://www.stat.berkeley.edu/users/terry/zarray/Html/index.html.
12. T. Ideker, V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Burngarner, D. R. Goodlett, R. Aebersold, and L. Hood, Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science*, **292**, pp. 929-934, 2001.
13. M. D. Ermolaeva, O. White, and S. L. Salzberg, Prediction of operons in microbial genomes, *Nucleic Acids Res.*, **29**, pp. 1216-1221, 2001.
14. L. A. Zadeh, Fuzzy sets, *Information and Control*, **8**, pp. 338-352, 1965.
15. J. M. Mendel, Fuzzy logic systems for engineering: a tutorial, *Proc. IEEE*, **83**, 345-377, 1995.
16. H. Zimmerman, *Fuzzy Set Theory and its Applications*, Kluwer, Boston, Massuchessets, 1999.
17. W. E. Combs and J. E. Andrews, Combinatorial rule explosion eliminated by a fuzzy rule configuration, *IEEE Trans. Fuzzy Syst.*, **6**, pp. 1-11, 1998.
18. L. A. Zadeh, Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets and Systems*, **1**, pp. 3-28, 1978.

19. P. J. Woolf and Y. Wang, A fuzzy logic approach to analyzing gene expression data, *Physiol. Genomics*, **3**, pp. 9-15, 2000.
20. W. E. Combs, Comment on Combinatorial rule explosion eliminated by a fuzzy rule configuration, Author s reply, *IEEE Trans. Fuzzy Syst.*, **7**, pp. 417-418, 1999.